©2017 Published in 5th International Symposium on Innovative Technologies in Engineering and Science 29-30 September 2017 (ISITES2017 Baku - Azerbaijan)

Determination of the Correlation between Air Quality and Diseases of Circulatory and Respiratory Systems

[1]Sultan Turhan and *[2]Defne Eskiocak
[1] Department of Computer Engineering, Galatasaray University, Turkey
*[2]Department of Computer Engineering, Galatasaray University, Turkey

Abstract

Compared with the other necessities of life, obligatory continuous consumption is a unique property of Air. Air pollution means that foreign matter in the air, which affects the health of humans and other living organisms negatively. In this study, we aimed to elaborate the correlation between air quality and Cardiovascular, Cerebrovascular and Respiratory Diseases. Air pollution is the contamination of the internal or external environment of any chemical, physical or biological agents and the modification of the natural characteristics of the atmosphere. The study is conducted in the province of Kirklareli, Turkey. Although there is no industrial activity in the region, air pollution rates are very high. With this project, we seek to determine the correlation between air quality and diseases of circulatory and respiratory systems. The health data is extracted Death Notification Systems of Turkish Republic's Ministry of Health, while the air quality data is provided by National Air Quality Monitoring Stations working under the auspices of Turkish Republic's Ministry of Environment and Urbanization. In the study, a data analysis using the linear regression method is realized and discovered a significant percentage of correlation between these two data sets.

Key words: Data analysis, air pollution, healthcare data, linear regression

1. Introduction

Especially in recent years, air pollution is a subject that attracts attention in the health sector. Air pollution is the contamination of the internal or external environment of any chemical, physical or biological agent and the modification of the natural characteristics of the atmosphere [1]. Pollutants that lead to serious health problems include particulate matter, carbon monoxide, nitrogen dioxide and sulfur dioxide [2]. One of the polluting substances is called "particulate matter", abbreviated "PM". PM is the term used for the mix of solid particles and liquid droplets in the air. These particles and droplets are extremely small and are made up of acids, organic chemicals, metals, soil particles and dust particles [3]. PM is taken in the body by inhalation due to their small size and after inhalation can lead to serious health problems [4]. According to the report published in 2014 by the World Health Organization, WHO [5], 7 million people worldwide died due to air pollution in 2012 [6].

As Anderson et al. (2012) have indicated; cardiovascular events and mortality rates are significantly higher in populations with long-term exposure to PM [3]. Dockery et al. (1993) observed 8111 people for 16 to 18 years and as a result of the study a 29% increase in the mortality rate was determined in the most polluted cities compared to the less polluted cities [7].

*Corresponding author: Address: Department of Computer Engineering, Galatasaray University, 34349, İstanbul TURKEY. E-mail address: defneeskiocak@gmail.com, Phone: +905358599799

Many previous studies have been based on air quality measurements, largely focusing on urban Pollution [8 –13]. Dominici et al (2006) estimate risks of cardiovascular and respiratory hospital admissions associated with short-term exposure to PM2.5 for Medicare enrollees and to explore heterogeneity of the variation of risks across regions and indicated that short-term exposure to PM2.5 increases the risk for hospital admission for cardiovascular and respiratory diseases [14]. Here we present results obtained the study conducted on the province of Kırklareli. Although there is no major industrial activity in the region air pollution rates are high. In order to reveal the correlation between the air quality and mortality caused by circulatory and respiratory systems' diseases, a data analysis using regression analysis will be presented.

The rest of the paper is organised as follows: in the next section, we describe the two data sets; their sources, structures and contexts. Then we explain the data preparation phase processed for each of them. After that we describe the data analysis phase and the linear regression model constructed. Finally, we close by summarising the results and suggestions for future research.

2. METHODS

2.1. Data Sources

To determine the correlation between air pollution and cardiovascular, cerebrovascular and respiratory diseases, two different data sets are collected. The first data set contains the data produced by National Air Quality Monitoring Stations [15] working under the auspices of Turkish Republic's Ministry of Environment and Urbanization [16]. The second one contains the data extracted from Death Notification Systems [17] of Turkish Republic's Ministry of Health [18].

1. National Air Quality Monitoring Stations' Data: These data are available on the web sites of Air Quality Monitoring Stations [15]. Thanks to these data it is possible to reach PM10 and sulfur dioxide (SO2) values of air pollution. The weather data are collected by monitoring stations via remote sensors located at various points of the region. Every day, hourly measurements are taken to determine the direction and speed of the wind, the temperature of the air, the air pressure and the relative humidity, the amount of rainfall, and the amounts of micrograms of components forming in one cubic meter of air.
2. Death Notification System's Data: This is the demographic data about the deaths and causes of deaths on the Death Notification System, managed by the Ministry of Health, the Turkish Public Health Authority. In Turkey, physicians working in local municipalities, primary care physicians, and community health center's practionners, or emergency physicians at health institutions may generate the death notification data. The data include the cause of death, contributing reasons, the definition of the disease causing the death and its relevant ICD10 [19] code with the demographic information obtained from the Central Population and Administration System (MERNIS) [20]. The records include;
    a. Unique number assigned to each individual
    b. City name where the death occurred
    c. District name where the death occurred
    d. Village name where the death occurred
    e. Institution name that registered the death

f. Sex of the individual
g. Profession of the individual
h. Nationality of the individual
i. Date of birth of the individual
j. Date of death of the individual
k. Age of the individual
l. Place of death
m. Reason of death (6 columns for main and underlying causes of death which are registered by hand)
n. Reason of death ICD-10 (4 columns for main and underlying causes of deaths that are selected from ICD-10 codes)
o. City of residence of the individual
p. District of residence of the individual

2.2. Data Processing

The study is conducted in the province of Kirklareli, Turkey. Although there is no industrial activity in the region, air pollution rates are very high. The region's air quality data are recorded since 2011 but in the study the data; which covers the period from 2013 till 2016 is used because the Death Notification System is operational only since 2013. Two data sets have different structures and contexts. Before determining the correlation between these two data sets, a cleaning and preparation process is conducted on the data.

2.2.1 Data Preparation Process for Death Notification System's Data

The data has been acquired in spreadsheet file format. The file consists of 7439 tuples. Each tuple corresponds to an individual death's records including the attributes cited above. For each tuple there are six attributes to specify the main and underlying causes of death as well as four attributes for underlying and underlying causes of death selected from ICD-10 codes. As a result of the primary analysis of the death cause attribute, some of the death causes such as drowning, car accident, injurry are found irrelevant to air pollution by their nature. So, another operation is realized under the supervision of Dr. Çiğdem Cerit, the Director of Kırklareli Public Health Provincial Directorate in order to determine the relevant death causes. We classify them into four categories that are "cardiovascular", "cerebrovascular", "respiratory" or "other" incidents and a new attribute called "disease category" is added to the data. While this operation, the tuples containing the disease name incorrectly written are also corrected.

After this cleaning operation, the data become ready to be analyzed. First, we consult the data to understand the structure and make interpretations by simple observation. Visualizing the data helps to better understand and interpret the data. Following attributes are chosen to interpret and visualize the data:

- Disease category
- District where the death occurred
- Age of the individual

- Sex of the individual

To learn more about our data, SQL queries are executed for each of the columns selected to study and the results are represented via the histograms. First, we try to determine the age interval of the people who are most affected from these disease groups. The results are represented in Figure 1.

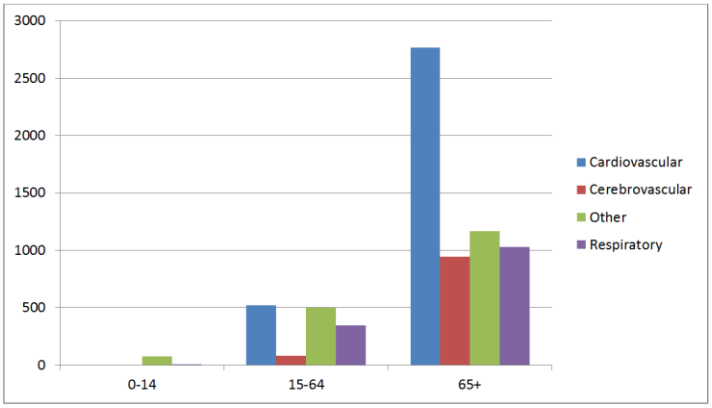

Figure 1. Number of deaths by age

Another query is executed to determine the mortality rate according to disease groups. The results are represented in Figure 2. The highest mortality rate belongs to deaths due to **cardiovascular** disease.

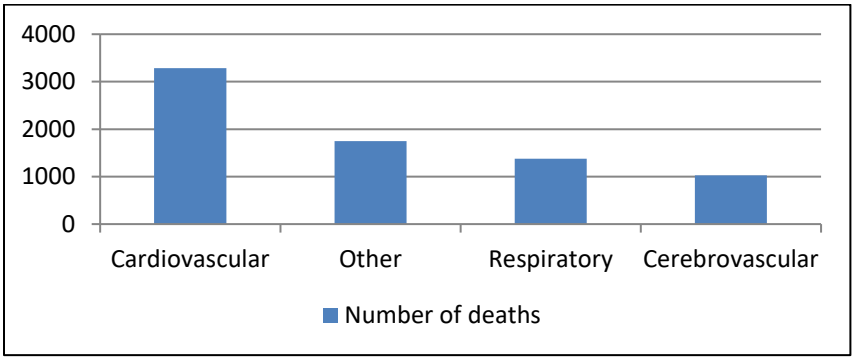

Figure 2. Number of deaths by disease category

The third query is lunched to determine death reason by disease category on the district and region basis. Figure 3 represents the distribution of diseases to the districts. The Districts "*Lüleburgaz*" and "*Merkez*" are affected the most by the deaths compared to the other districts

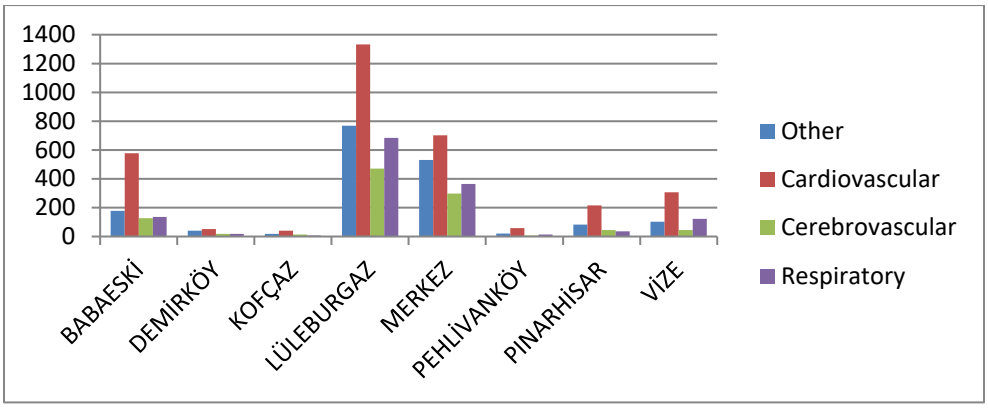

Figure 3. Distribution of diseases in the districts

2.2.2 Data Preparation Process for Air Quality Data

As well as health data, a data preparation process is realized for the air quality data measured in Kırklareli. Four districts of Kırklareli; Merkez, Lüleburgaz, Demirköy (also known as "Limanköy") and Vize have air quality monitoring stations. The data produced by these stations is published publicly on National Air Quality Monitoring Network's website and can be exported as spreadsheet format. For the study, the records between January 1, 2013 and December 31, 2016 are only taken. As air quality is measured with several parameters on hourly basis, the extracted data contains 35074 tuples on average for a district. According to their capacity, the air quality station of each district measures different sets of parameters. In order to provide the homogeneity between them, we decide to include the parameters PM_{10}, SO_2, NO_2, and O_3 into the study. To qualify the air, Turkish Republic's Ministry of Environment and Urbanization uses the values shown in Table 1 for these parameters. Although the air quality index accepted by WHO differs from the values accepted by Ministry of Environment and Urbanization, we will use the latter for our study.

Table 1. Air Quality Index

Index	AQI	SO_2	PM_{10}	NO_2	O_3
Good	0-50	0-100	0-50	0-100	0-120
Average	51-100	101-250	51-100	101-200	121-160
Delicate	101-150	251-500	101-260	201-500	161-180
Unhealthy	151-200	501-850	261-400	501-1000	181-240
Poor	201-300	851-1100	401-520	1001-2000	241-700
Dangerous	301-500	>1100	>521	>2001	>701

The primary inspection of air quality data reveals that there are too many null values in Vize's and Demirköy's data set due to fact that these air quality measurement stations don't work at those dates. For example, in Vize, the first measurement was made on September 24, 2014; 21 months later than the beginning of our research period. We first tried to derive and produce the data from existing ones and fulfill them. Unfortunately the first analysis bring out that the synthetic data produced in such a way has any significant impact in neither positive nor negative

way on the results. We then decided to remove these two districts from the study and concentrate on Merkez's and Lüleburgaz's data that contain any null or damaged values. The second issue that influences our decision to put Vize and Demirköy out of the study is that the number of deaths in these districts is significantly low, when compared to the others. This situation reduces the accuracy and effectiveness of the analysis. The numbers of deaths according to districts are given in Table 2.

Table 2. Number of deaths in districts

District	Cardiovascular	Cerebrovascular	Respiratory	Other	% of the population
Merkez	703	298	364	532	27,76%
Lüleburgaz	1333	472	684	769	41,30%
Demirköy	52	18	18	41	2,41%
Vize	306	45	122	103	7,84%

2.3. Data Analysis

2.3.1 The Analysis Method

To determine the correlation between these two data sets that are completely different each other, we decided to use *Linear Regression* method [21]. This method focuses on the form of the relationship between variables, while the objective of correlation analysis is to gain insight into the strength of the relationship [21] It is an approach to modeling the association between a numeric dependent variable *Y* and one or more independent variables denoted *X*. The case of one explanatory variable in regression model is called simple linear regression. For more than one explanatory variable, then the model is called multiple linear regressions. The dependent variable should be a numeric variable in linear regression. However, it should be noted that a statistically significant regression analysis does not imply the causal relationship between the two sets of data that it reveals to be correlated. The existing causal relationship relation is often interpreted by a lurking variable that is not found in both data sets [22].

2.3.2 The Model

In the study, the air quality data set is modeled as independent variable and the number of deaths as dependent variable. In order to determine the relevant data, first we identify the dates when the pollutants are above the safe limits determined by Ministry of Environment and Urbanization. The air quality index represented in Table 1 is used to determine the air nature for each date and only the dates where the air has the index "delicate", "unhealthy", "poor" or "dangerous" are taken into consideration to create the working data set. To create the model's data set, we consider the date where the pollutants are above the safe limits as the initial date t_0 and execute a SQL query to calculate number of deaths by districts for this date. The second SQL query is responsible for calculating cumulated number of deaths by districts between this date and a month before t_0-1. The third one is responsible for finding cumulated number of deaths by districts between this date and a month after this date t_0+1. To automate this process, a batch program is developed in Java. Once the data set is created, the simple linear regression analysis is

performed with Weka [23] which is a modern software platform for applied machine learning.

2.3.2 Results

The first analysis is realized on Weka platform for Merkez district. The corresponding simple linear regression model is represented in Figure 3. The system find out a correlation coefficient of 0,8386 which is a significant value.

```
Posayisi =

     -0      * tarih +
     -0.459  * severity +
      0.0558 * Mosayisi +
      6.7351 * onedeni=respiratoire,autre,cardiovasculaire +
      4.9562 * onedeni=autre,cardiovasculaire +
     13.4738 * onedeni=cardiovasculaire +
    136.5692

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8386
Mean absolute error                      4.8144
Root mean squared error                  6.5837
Relative absolute error                 52.7196 %
Root relative squared error             54.4192 %
Total Number of Instances            2376
```

Figure 3. Merkez

The second analysis is realized on Weka platform for Lüleburgaz district. The corresponding simple linear regression model is represented in Figure 4. The system find out a correlation coefficient of 0,877 which is a significant value representing a closer correlation.

```
Posayisi =

      0      * tarih +
      0.2394 * Mosayisi +
     10.2466 * onedeni=respiratoire,autre,cardiovasculaire +
      4.1543 * onedeni=autre,cardiovasculaire +
     25.4768 * onedeni=cardiovasculaire +
   -199.7625

Time taken to build model: 0.02 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.877
Mean absolute error                      8.9864
Root mean squared error                 11.3355
Relative absolute error                 43.1797 %
Root relative squared error             48.0436 %
Total Number of Instances            18268
```

Figure 4. Lüleburgaz

When interpreting the results of a simple linear regression analysis, it is not sufficient to take consider only the correlation coefficient. The root mean squared error also should be taken into account. The lower is this value, the more accurate the regression line drawn for the data set is. In this case, when both results are observed, it is found out that the root mean squared error values is quite low. Therefore, it is possible to conclude that the models are optimally designed.

3. Discussion

In this study, a correlation between air quality and circulatory and respiratory systems diseases is discovered upon the results of simple linear regression analysis for two Kırklareli's districts. Defining the correlation is the first step to achieve in the study since it cannot explain the causal relationship despite the analysis reveals the existence of the correlation between them. For this reason, in the near future the study will continue by further analysis through using complex network analysis in order to reveal the lurking variable to construe the causal relationship between them.

4. Conclusions

The objective of this study was to determine if there is a correlation between air quality and circulatory and respiratory systems' diseases. The analysis is realized with the health data acquired from the "Death Notification System" and the air quality data collected through Ministry of Environment and Urban Planning. The analysis is limited to Kırklareli and its province and the dates between years 2013 and 2016. We used the simple linear regression method to analyse the collected data and discovered a significant correlation between air quality and aforementioned diseases.

The results of the analysis may help to take concrete steps on the optimization of the services provided by the Ministry of Health throughout the region. Effective and efficient outcomes have been achieved for both patient and healthcare providers, for example, in determining the most suitable locations for opening specific services such as the coronary intensive care unit, and determining the most efficient way of employing specialists such as cardiologists. At the same time, reducing the costs of the necessary treatments is intended by decreasing the possibility of future diseases by determining the places where the environmental health service should be provided and the types of preventive services that are needed.

Acknowledgements

This study is carried out with Galatasaray University Ethics Committee's permit dated 20/01/2017 and no. 40188. It is supported financially by Galatasaray University Scientific Research Project no: 15.401.003. The authors would like to thank Kirklareli Public Health Director Dr. Cigdem Cerit, Kirklareli Infectious Diseases, Environment and Employee Health Branch Manager Dr. Aycin Ugur and Galatasaray University Scientific Research Council for their valuable contribution to the study.

References

[1] Goldsmith, J. R., & Friberg, L. T. (1977). Effects of air pollution on human health. *Air pollution*, *2*, 457-610.
[2] Wark, K., & Warner, C. F. (1981). Air pollution: its origin and control.
[3] Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: a review of the effects of particulate matter air pollution on human health. Journal of Medical Toxicology. 2012 Jun 1;8(2):166-75.
[4] US Environmental Protection Agency PM Pollution. https://www.epa.gov/pm-pollution. Accessed: 2017-07-24.
[5] World Health Organization, http://www.who.int
[6] World Health Organization Deaths by Air Pollution in 2014 Report. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/. Accessed: 2017-07-24.
[7] Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., ... & Speizer, F. E. (1993). An association between air pollution and mortality in six US cities. *New England journal of medicine*, 329(24), 1753-1759.
[8] Ostro, B. Outdoor Air Pollution: Assessing the Environmental Burden of Disease at National and Local Levels (World Health Organization Environmental Burden of Disease Series No. 5, WHO, Geneva, 2004)
[9] Cohen, A. J. et al. The global burden of disease due to outdoor air pollution. J. Toxicol. *Environ. Health A* 68, 1301–1307 (2005)
[10] Pope, C. A., III et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J. Am. Med. Assoc*. 287, 1132–1141 (2002)
[11] Prüss-Üstün, A., Bonjour, S. & Corvalan, C. The impact of the environment on health by country: a meta-synthesis. *Environ. Health 7*, http://dx.doi.org/10.1186/1476-069X-7-7 (2008)
[12]Russell, A. G. & Brunekreef, B. A focus on particulate matter and health. *Environ. Sci. Technol.* 43, 4620–4625 (2009)
[13] Gurjar, B. R. et al. Human health risks in megacities due to air pollution. *Atmos. Environ.* 44, 4606–4613 (2010)
[14] Dominici, F., Peng, R.D., Bell, M.L., Pham, L., McDermott, A., Zeger, S.L. and Samet, J.M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, Volume 295, Issue 10, pp.1127-1134
[15] Ulusal Hava Kalitesi İzleme Ağı http://www.havaizleme.gov.tr/Default.ltr.aspx Accessed 2017 – 07 – 24
[16] T.C. Çevre ve Şehircilik Bakanlığı https://www.csb.gov.tr Accessed 2017 – 07 – 24
[17] Türkiye Halk Sağlığı Kurumu Ölüm Bildirim Sistemi https://obs.gov.tr/ Accessed 2017 – 07 – 24
[18] Türkiye Cumhuriyeti Sağlık Bakanlığı https://www.saglik.gov.tr/ Accessed 2017 – 07 – 24
[19] ICD – Internation Classification of Disease http://www.who.int/classifications/icd/en/ Accessed 2017 – 07 – 24
[20] Merkesi Nüfus İdare Sistemi https://www.nvi.gov.tr Accessed 2017 – 07 – 24
[21] Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3), 617-628.

[22] Dawson B, Trapp RG. Statistical Methods for Multiple Variables. Basic & Clinical Biostatistics. Lange Medical books/McGraw Hill Medical Publishing Division, 2001, USA, 236-242

[23] WEKA http://www.cs.waikato.ac.nz/ml/weka/ Accessed 2017- 07-24